Report

Shriya Singaraju (ss3886), Anusha Paranjpe (ap1993), Julia Hansen (jlh463), Jasmine Hanjra (jkh122), Tanvisri Munagala (tm900)

2023-04-28

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(moderndive)
library(ggplot2)
library(skimr)
library(infer)
```

Exploratory data analysis:

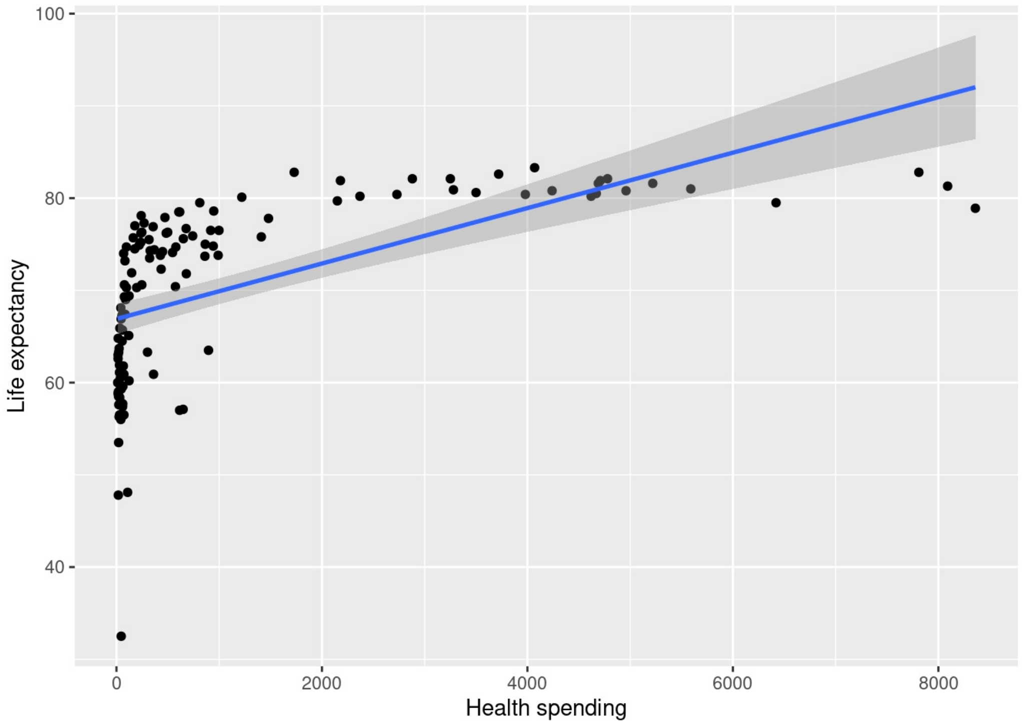
```
data <- read.csv("HW1.csv", header = T)
```

Plot 1: Life expectancy vs. Health spending

The plot shows a positive relationship between health spending and life expectancy. Countries with higher health spending tend to have higher life expectancy. However, the relationship is not perfect, as there are some countries with lower health spending that have higher life expectancy than some countries with higher health spending.

```
# Plot 1: Life expectancy vs. Health spending
ggplot(data, aes(x = healthspend, y = lifeexp)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Health spending", y = "Life expectancy")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

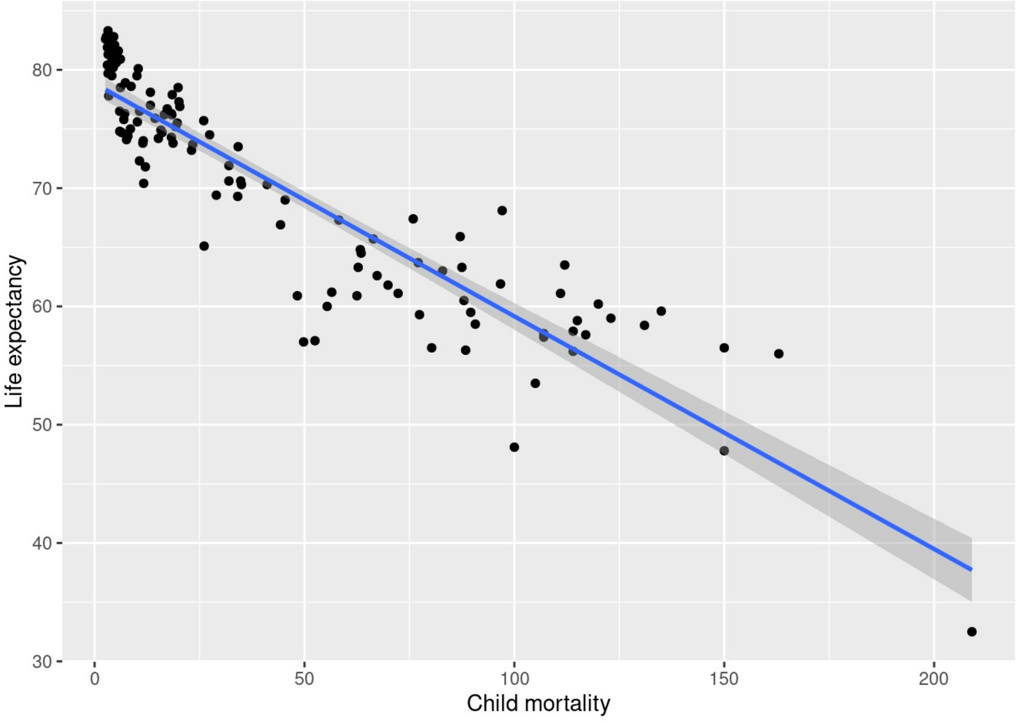


Plot 2: Life expectancy vs. Child mortality

The plot shows a negative relationship between child mortality and life expectancy. Countries with lower child mortality tend to have higher life expectancy. The relationship is strong, as the data points are tightly clustered around the trend line.

```
# Plot 2: Life expectancy vs. Child mortality
ggplot(data, aes(x = childmort, y = lifeexp)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Child mortality", y = "Life expectancy")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

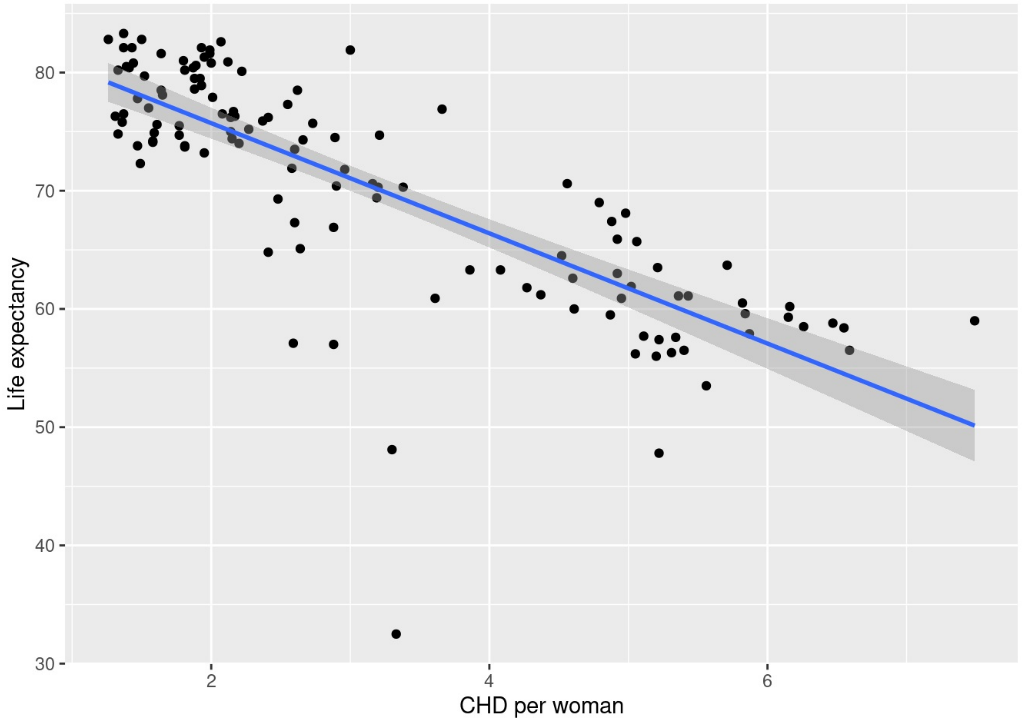


Plot 3:Life expectancy vs. CHD per woman

The plot shows a negative relationship between CHD (coronary heart disease) per woman and life expectancy. Countries with higher CHD per woman tend to have lower life expectancy. However, there is some variation in the data, as some countries with higher CHD per woman have higher life expectancy than some countries with lower CHD per woman.

```
# Plot 3: Life expectancy vs. CHD per woman
ggplot(data, aes(x = chdperwoman, y = lifeexp)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "CHD per woman", y = "Life expectancy")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

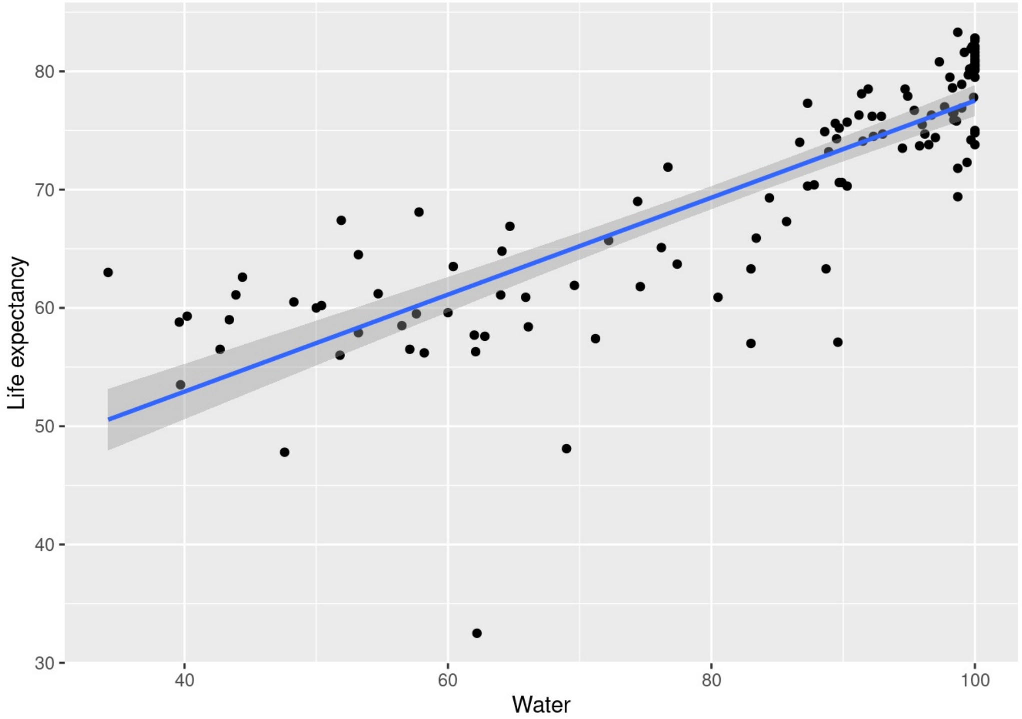


Plot 4: Life expectancy vs. Water

The plot shows a positive relationship between access to clean water and life expectancy. Countries with higher access to clean water tend to have higher life expectancy. However, there is some variation in the data, as some countries with lower access to clean water have higher life expectancy than some countries with higher access to clean water.

```
# Plot 4: Life expectancy vs. Water
ggplot(data, aes(x = water, y = lifeexp)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Water", y = "Life expectancy")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

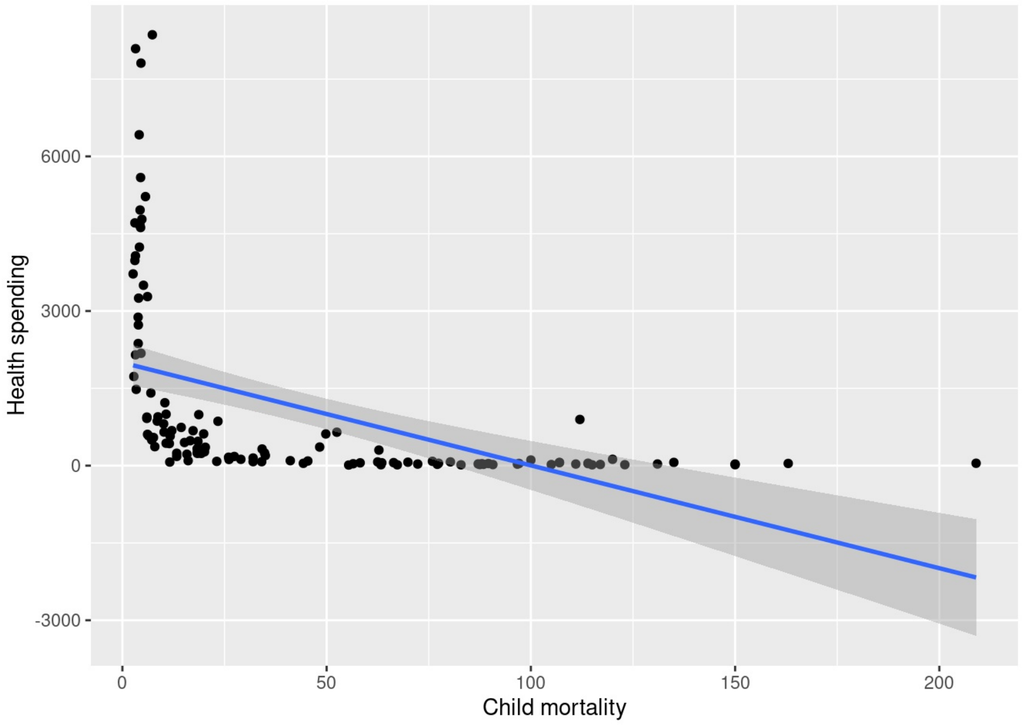


Plot 5: Health spending vs. Child mortality

The plot shows a negative relationship between health spending and child mortality. Countries with higher health spending tend to have lower child mortality. However, there is some variation in the data, as some countries with lower health spending have lower child mortality than some countries with higher health spending.

```
# Plot 5: Health spending vs. Child mortality
ggplot(data, aes(x = childmort, y = healthspend)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Child mortality", y = "Health spending")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

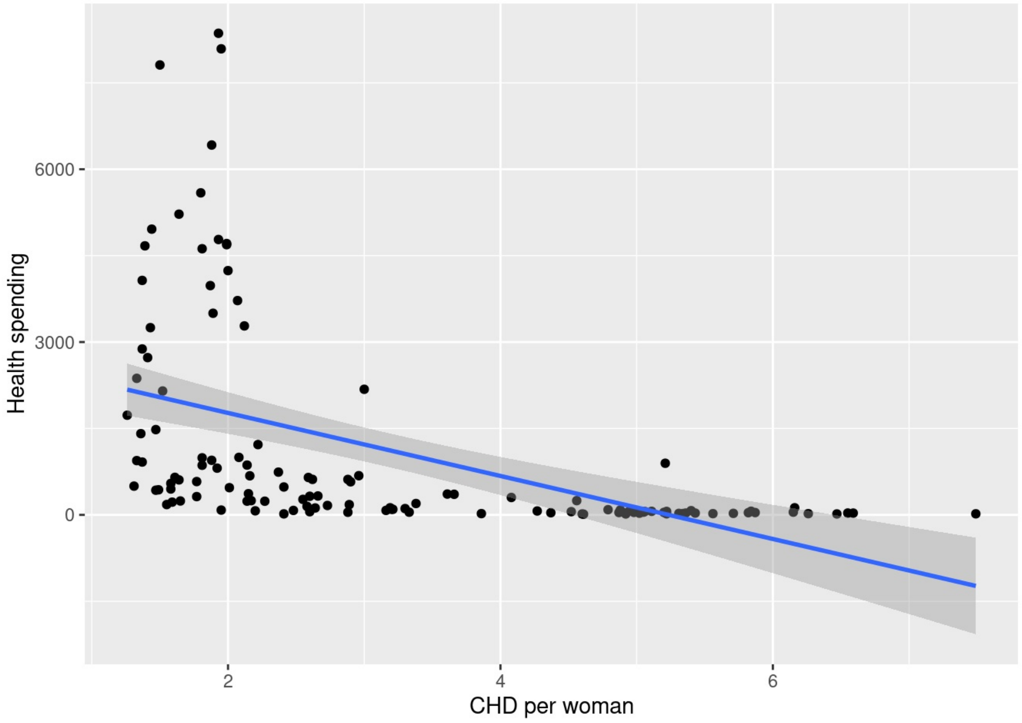


Plot 6: Health spending vs. CHD per woman

The plot shows a weak negative relationship between health spending and CHD per woman. Countries with higher health spending tend to have slightly lower CHD per woman, but the relationship is not very strong.

```
# Plot 6: Health spending vs. CHD per woman
ggplot(data, aes(x = chdperwoman, y = healthspend)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "CHD per woman", y = "Health spending")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

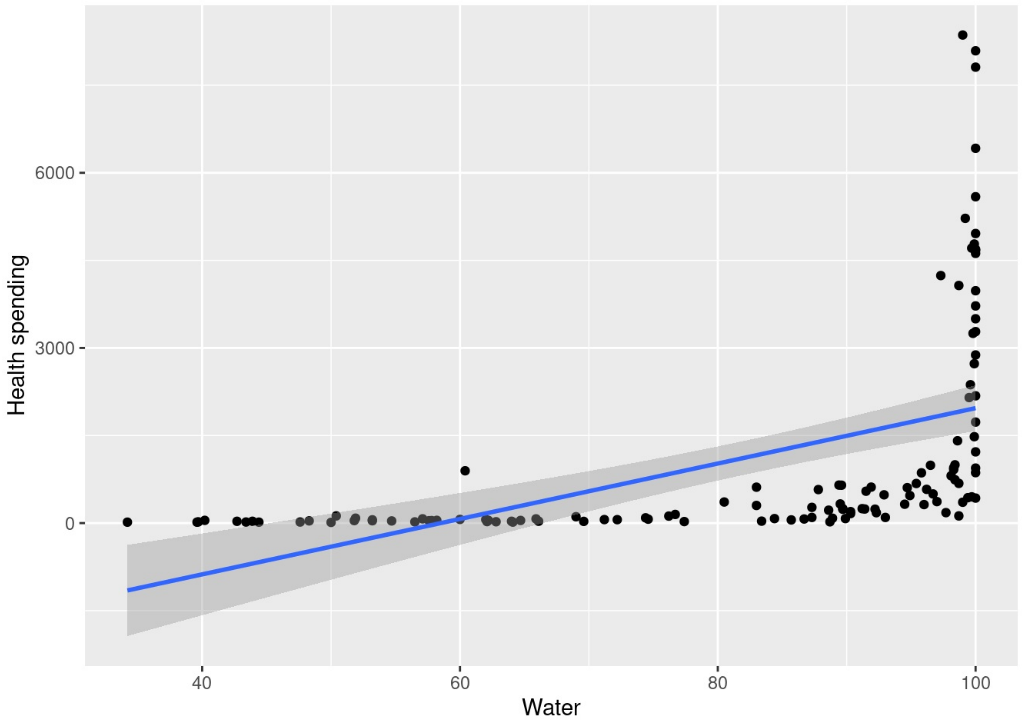


Plot 7: Health spending vs. Water

The plot shows a positive relationship between health spending and access to clean water. Countries with higher health spending tend to have higher access to clean water. However, the relationship is not perfect, as there are some countries with lower health spending that have higher access to clean water than some countries with higher health spending.

```
# Plot 7: Health spending vs. Water
ggplot(data, aes(x = water, y = healthspend)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Water", y = "Health spending")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

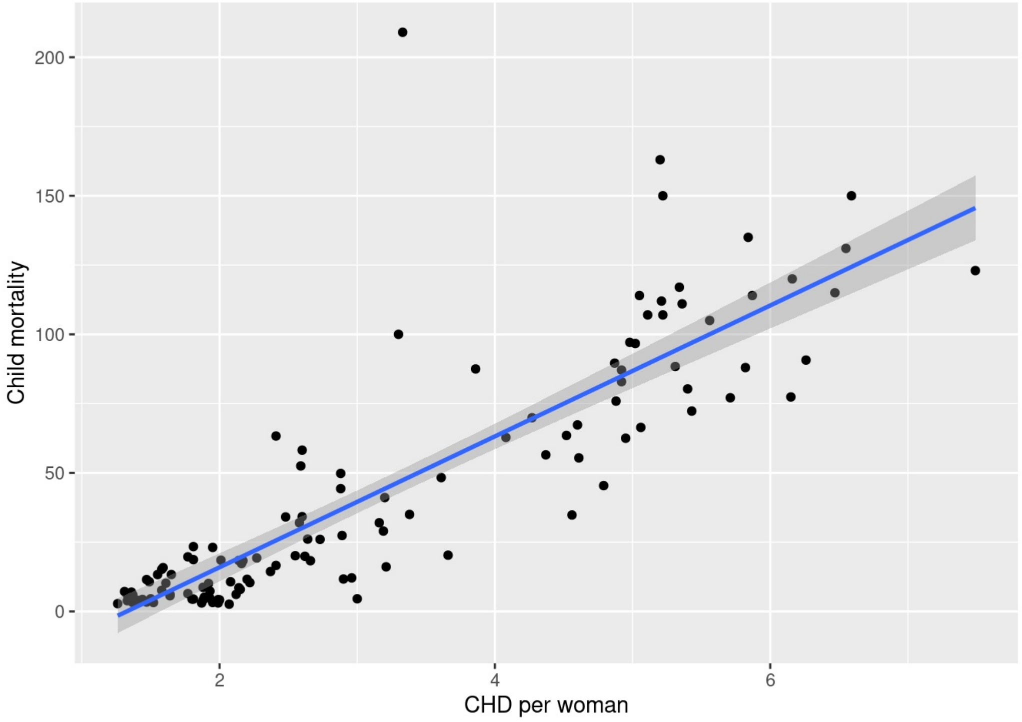


Plot 8: Child mortality vs. CHD per woman

The plot shows a weak positive relationship between child mortality and CHD per woman. Countries with higher child mortality tend to have slightly higher CHD per woman, but the relationship is not very strong.

```
# Plot 8: Child mortality vs. CHD per woman
ggplot(data, aes(x = chdperwoman, y = childmort)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "CHD per woman", y = "Child mortality")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

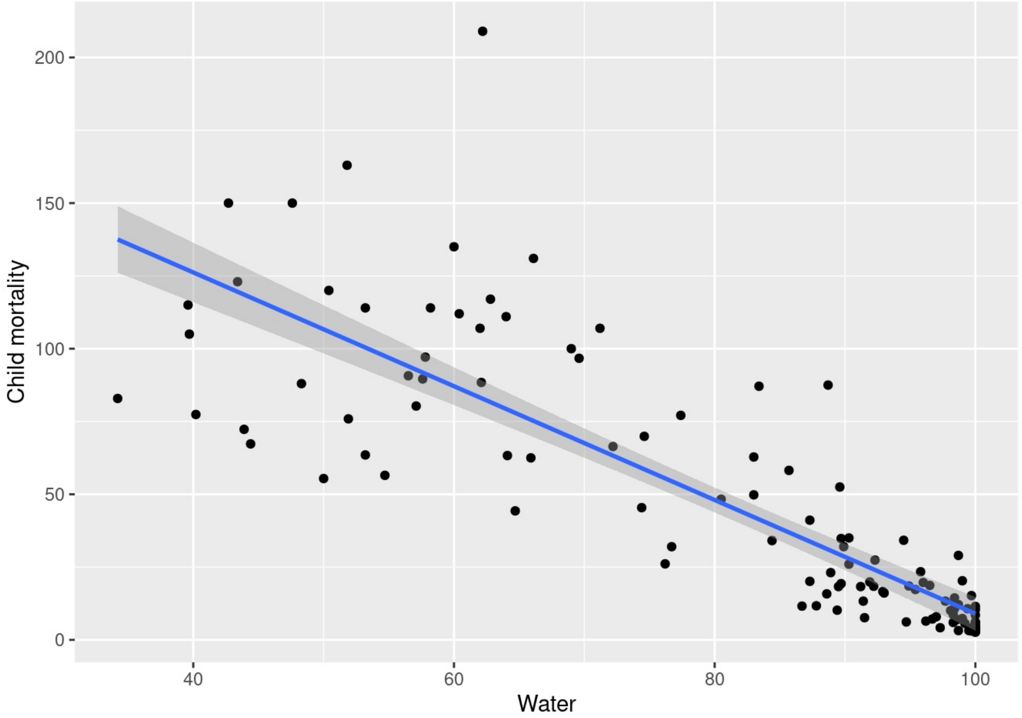


Plot 9: Child mortality vs. Water

The plot shows a negative relationship between access to clean water and child mortality. Countries with higher access to clean water tend to have lower child mortality. However, the relationship is not perfect, as there are some countries with lower access to clean water that have lower child mortality than some countries with higher access to clean water.

```
# Plot 9: Child mortality vs. Water
ggplot(data, aes(x = water, y = childmort)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Water", y = "Child mortality")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

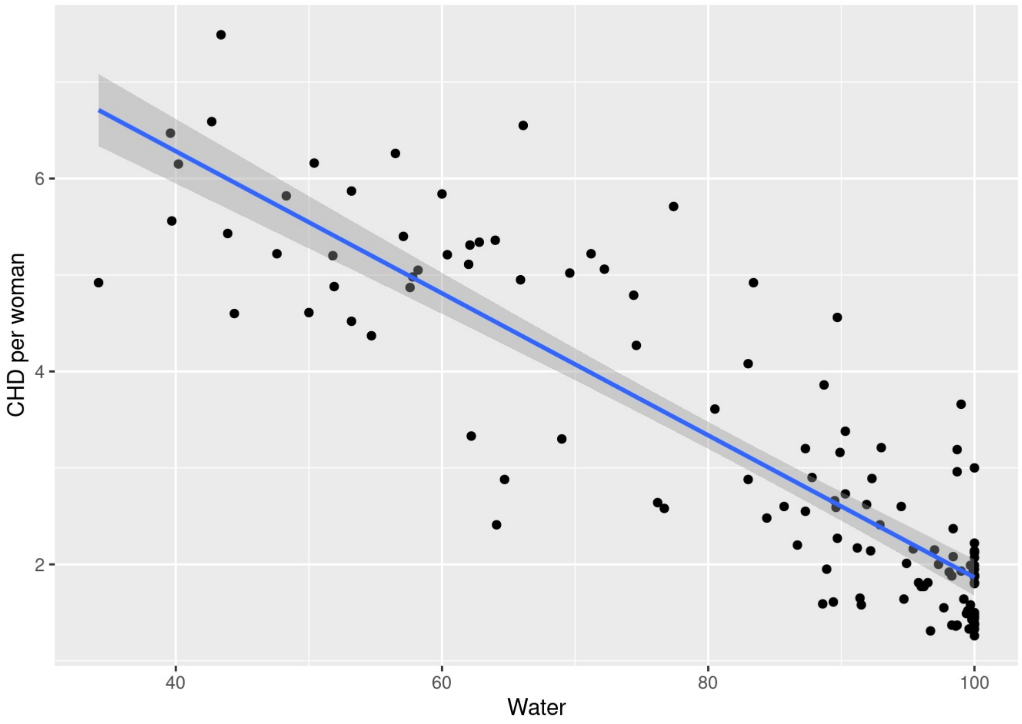


Plot 10: CHD per woman vs. Water

The plot shows a negative relationship between CHD per woman and access to clean water. Countries with higher access to clean water tend to have lower CHD per woman. However, the relationship is not perfect, as there are some countries with lower access to clean water that have lower CHD per woman than some countries with higher access to clean water.

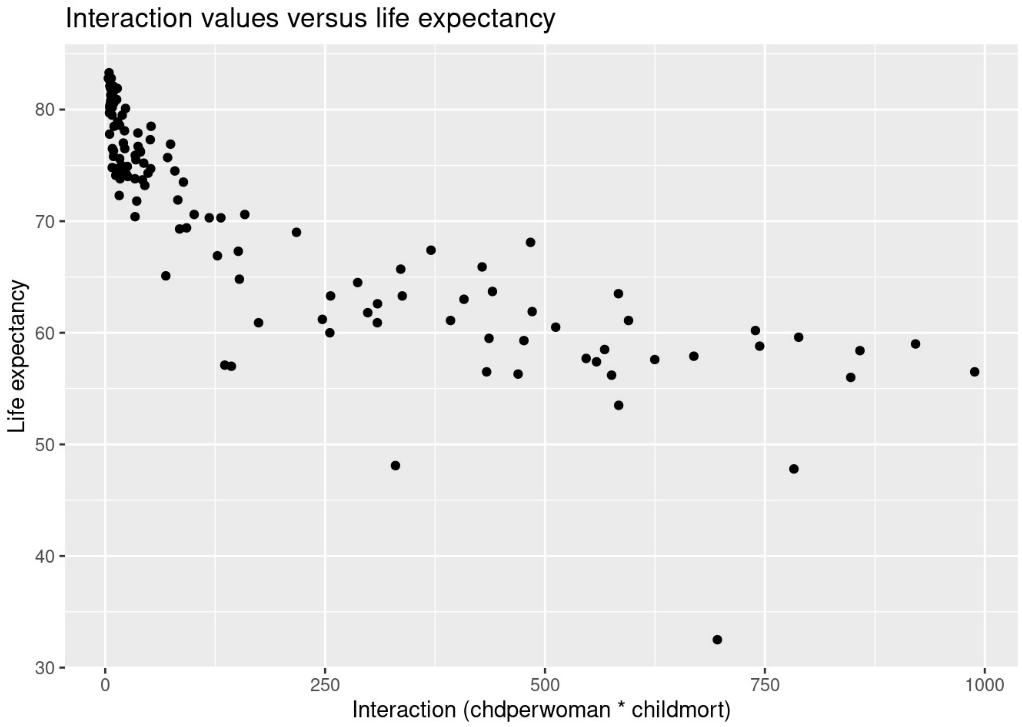
```
# Plot 10: CHD per woman vs. Water
ggplot(data, aes(x = water, y = chdperwoman)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Water", y = "CHD per woman")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Interaction model plot

```
ggplot(data, aes(x = chdperwoman * childmort, y = lifeexp)) +
  geom_point() +
  xlab("Interaction (chdperwoman * childmort)") +
  ylab("Life expectancy") +
  ggtitle("Interaction values versus life expectancy")
```

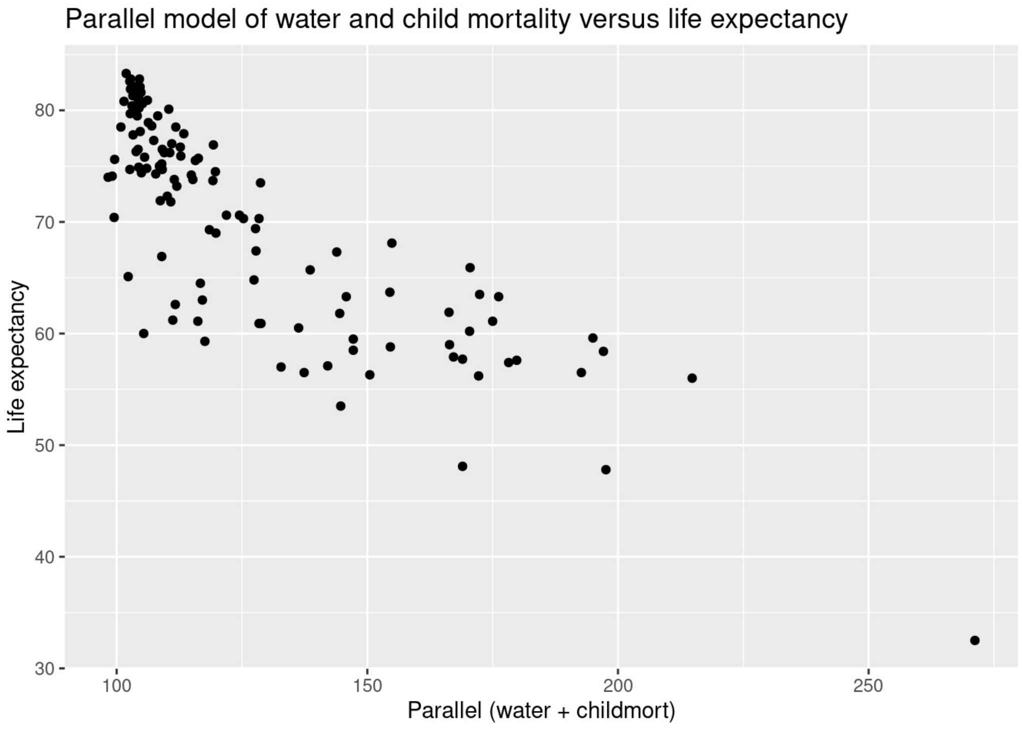


The plot shows that there is a negative relationship between the interaction of childbirths and child mortality and life expectancy. Countries with higher life expectancy tend to have lower values of the interaction, while countries with lower life expectancy tend to have higher values of the interaction. This suggests that the interaction of childbirths and child mortality is a factor that contributes to lower life expectancy in some countries.

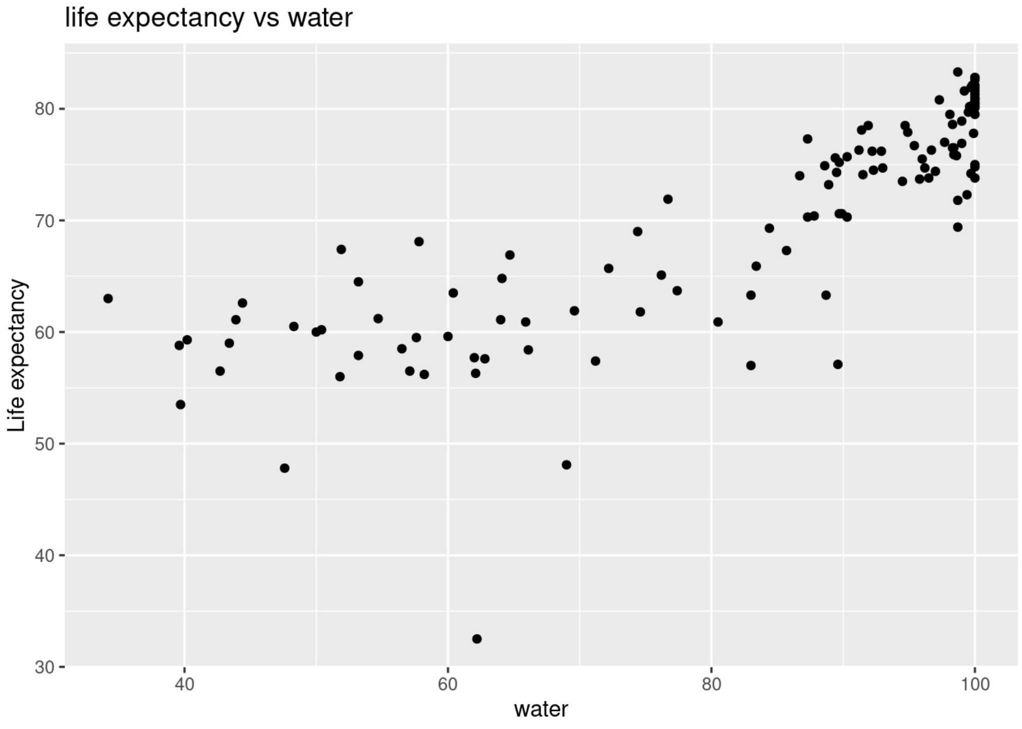
Further Analsyis

parallel model plot

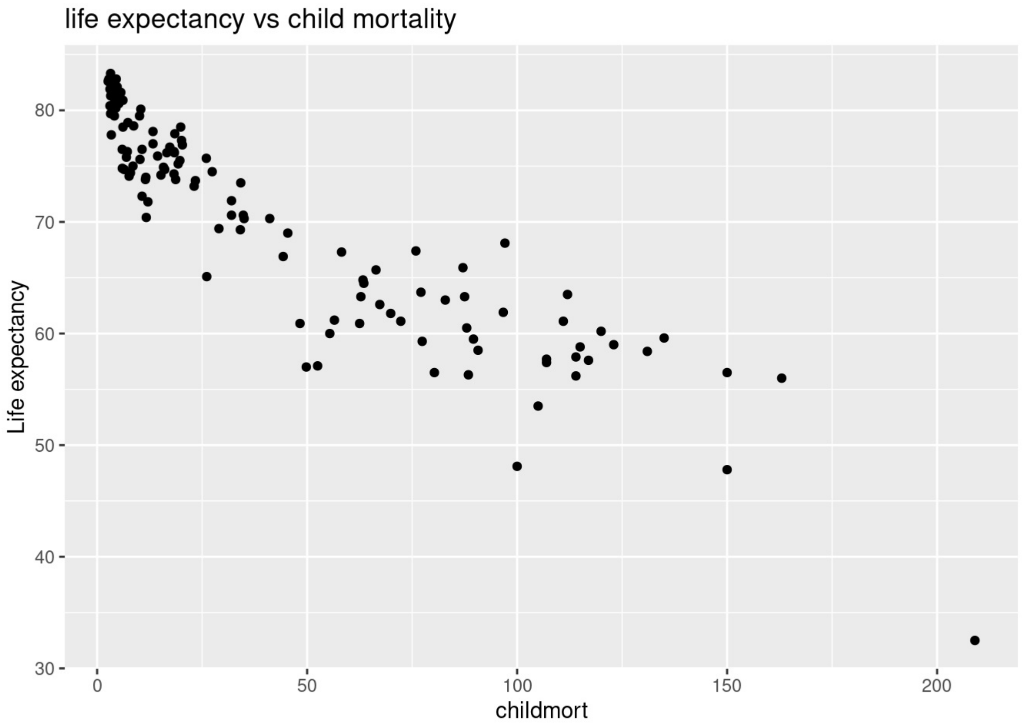
```
ggplot(data, aes(x = water + childmort, y = lifeexp)) +
  geom_point() +
  xlab("Parallel (water + childmort)") +
  ylab("Life expectancy") +
  ggtitle("Parallel model of water and child mortality versus life expectancy")
```



```
# separate analysis of lifeexp vs. water
ggplot(data, aes(x = water, y = lifeexp)) +
  geom_point() +
  xlab("water") +
  ylab("Life expectancy") +
  ggtitle("life expectancy vs water")
```



```
# separate analysis of lifeexp vs. child mortality
ggplot(data, aes(x = childmort, y = lifeexp)) +
  geom_point() +
  xlab("childmort") +
  ylab("Life expectancy") +
  ggtitle("life expectancy vs child mortality")
```


life expectancy vs child mortality

The plot shows that there is a negative relationship between the parallel model of water and child mortality as the regressors, and life expectancy. This suggests that there is a variable that contributes to the increase in life expectancy as the regressor value increases. There also appears to be an outlier in the data where life expectancy is between 30-35 years old. Separate analysis of water versus life expectancy reveals a slight positive relationship when the percentage of people who have access to water increases beyond 80%, because life expectancy increases . However, below 80, there seems to be no discernable relationship between increase in access to water and life expectancy. An individual analysis of child mortality versus life expectancy shows a negative relationship between the two variables, suggesting that an increase in life expectancy is correlated with a decrease in child mortality.

Primary Final Model: Model for Predictions:

In order to determine our final model we tested each of the variables against the others to find the model with the optimized adjusted R^2 and fit the LINE assumptions.

We first created a correlation matrix in order to take note of the relationship between the variables in our table and also to keep track of the correlation between our dependent variables.

```
temp <- select(data, population, childmort, income, gdpcapita, chdperwoman, healthspend, co2, water, popdensity,
murder)
cor(temp)
```

```
##               population    childmort      income   gdpcapita chdperwoman
## population   1.000000000 -0.031440438 -0.07155391 -0.05090997 -0.10389057
## childmort   -0.031440438  1.000000000 -0.63088658 -0.53310833  0.85910987
## income      -0.071553906 -0.630886580  1.00000000  0.93544396 -0.62746441
## gdpcapita   -0.050909966 -0.533108329  0.93544396  1.00000000 -0.52080820
## chdperwoman -0.103890572  0.859109874 -0.62746441 -0.52080820  1.00000000
## healthspend -0.020499743 -0.478182728  0.83977983  0.94773773 -0.47704677
## co2         -0.007336614 -0.507671392  0.74276584  0.61220365 -0.50737932
## water        0.046928750 -0.848412162  0.66281511  0.55432074 -0.87908315
## popdensity   0.001758586 -0.110569550  0.32749296  0.20475354 -0.14590716
## murder       0.579662474  0.007810528 -0.11245282 -0.09724403 -0.06654685
##             healthspend          co2       water   popdensity       murder
## population  -0.02049974 -0.007336614  0.04692875  0.001758586  0.579662474
## childmort   -0.47818273 -0.507671392 -0.84841216 -0.110569550  0.007810528
## income       0.83977983  0.742765844  0.66281511  0.327492963 -0.112452822
## gdpcapita    0.94773773  0.612203650  0.55432074  0.204753542 -0.097244033
## chdperwoman -0.47704677 -0.507379320 -0.87908315 -0.145907164 -0.066546846
## healthspend  1.00000000  0.503621853  0.49423735  0.036664746 -0.048223317
## co2          0.50362185  1.000000000  0.54421009  0.158974242 -0.066163219
## water        0.49423735  0.544210091  1.00000000  0.125109132  0.051850296
## popdensity   0.03666475  0.158974242  0.12510913  1.000000000 -0.023552559
## murder      -0.04822332 -0.066163219  0.05185030 -0.023552559  1.000000000
```

```
#individual regression of lifeexp~childmort
model_childmort <- lm(data=data, lifeexp ~ childmort)
get_regression_table(model_childmort)
```

```
## # A tibble: 2 × 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   78.8       0.493     160.        0   77.9     79.8
## 2 childmort   -0.197     0.008     -24.8       0   -0.212   -0.181
```

```
#Our p-values are both 0, which indicates that the Beta coefficient is significant
get_regression_summaries(model_childmort)
```

```
## # A tibble: 1 × 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.836         0.834  15.1  3.89  3.92      615.       0     1   123
```

Among the variables we tested it was immediately noticed that the R^2 value of this model of child mortality on life expectancy was 0.836 and the adjusted_r_squared value is 0.834. This indicated to us that that child mortality would be a good basis for us to build our final model on. Using this we continued to develop models based on the interaction between select variables. Noting the r_squared and adjusted_r_squared values we continued to add and subtract variables in order to optimize the r_squared and aj_r_squared of our model. Following is the code for 3 significant models that lead us to our final model:

```
#interaction model chdperwoman and childmort
model_chdwoman <- lm(data=data, lifeexp ~ chdperwoman * childmort)
get_regression_table(model_chdwoman)
```

```
## # A tibble: 4 × 7
##   term                  estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                    <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept               84.5       1.07      78.9        0   82.4     86.6
## 2 chdperwoman             -2.24      0.461     -4.85       0   -3.15    -1.32
## 3 childmort               -0.355     0.025    -14.1        0   -0.405   -0.305
## 4 chdperwoman:childmort    0.041     0.006      7.43       0    0.03     0.052
```

```
#All the p-values are significant for each of the Beta coefficients in this model


#adding healthspend
model_4a <- lm(data=data, lifeexp ~ healthspend + (chdperwoman * childmort))
get_regression_table(model_4a)
```

```
## # A tibble: 5 × 7
##   term                  estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                    <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept               82.2       1.27      64.6    0       79.7     84.7
## 2 healthspend              0.001     0          3.13   0.002    0        0.001
## 3 chdperwoman             -1.74      0.473     -3.67   0       -2.67    -0.8
## 4 childmort               -0.324     0.026    -12.4    0       -0.376   -0.273
## 5 chdperwoman:childmort    0.035     0.006      6.13   0        0.024    0.047
```

```
#significant, beta coefficients for healthspend + (chdperwoman * childmort)
get_regression_summaries(model_4a)
```

```
## # A tibble: 1 × 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.896         0.893  9.54  3.09  3.15      255.       0     4   123
```

```
#0.896 r_squared value and adjusted_r_squared value of 0.893

pop <- as.factor(ifelse(data$population<1.090e+07, 1,0))
df2 <- cbind(data, pop)
model_01 <- lm(data=data, lifeexp ~ healthspend + (chdperwoman * childmort) + water + pop)
get_regression_table(model_01)
```

```
## # A tibble: 7 × 7
##   term                  estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                    <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept               71.4       3.80     18.8     0       63.8     78.9  
## 2 healthspend              0.001     0         2.87    0.005    0        0.001
## 3 chdperwoman             -1.05      0.513    -2.05    0.043   -2.07    -0.036
## 4 childmort               -0.307     0.026   -11.7     0       -0.359   -0.255
## 5 water                    0.099     0.033     3.01    0.003    0.034    0.165
## 6 pop: 1                  -0.12      0.565    -0.213   0.832   -1.24     0.998
## 7 chdperwoman:childmort    0.035     0.006     6.17    0        0.023    0.046
```

The pop variable’s beta coefficient has a p-value of 0.832, not significant, so this indicates that the Beta coefficient is not a non-zero value.

```
get_regression_summaries(model_01)
```

```
## # A tibble: 1 × 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.904         0.899  8.84  2.97  3.06      182.       0     6   123
```

While the r_squared value is also 0.904 we ultimately discarded this model as the adjusted r_squared value dropped to 0.899, indicating that the population variable did not add to overall accuracy of the model.

From this process we developed our final model to be:

```
model_chosen <- lm(data=data, lifeexp ~ healthspend + (chdperwoman * childmort) + water)
get_regression_table(model_chosen)
```

```
## # A tibble: 6 × 7
##   term                  estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                    <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept               71.4       3.79      18.9    0       63.9     78.9  
## 2 healthspend              0.001     0          2.88   0.005    0        0.001
## 3 chdperwoman             -1.05      0.511     -2.06   0.042   -2.06    -0.04 
## 4 childmort               -0.308     0.026    -11.9    0       -0.359   -0.257
## 5 water                    0.098     0.033      3.02   0.003    0.034    0.163
## 6 chdperwoman:childmort    0.035     0.006      6.23   0        0.024    0.046
```

```
get_regression_summaries(model_chosen)
```

```
## # A tibble: 1 × 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.904           0.9  8.85  2.97  3.05      220.       0     5   123
```

The Effect of the Predictor Variables on the Life Expectancy.

y = 71.4 + 0.001(x1) - 1.05(x2) - 0.308(x3) + 0.098(x4) + 0.035(x2)*(x3)

The expected increase in life expectancy when all independent variables are 0 is 71.4.

The expected increase in life expectancy when health is increased by 1 unit and all other variables are held constant is 0.001.

The expected increase in life expectancy when chdperwoman increased by 1 unit and all other variables are held constant is -1.05 + 0.006*(x3), where x3 = childmort.

The expected increase in life expectancy when childmort increased by 1 unit and all other variables are held constant is -0.308 + 0.006*(x2), where x2 = chdperwoman.

The expected increase in life expectancy when water is increased by 1 unit and all other variables are held constant is 0.098.

We ultimately chose this model because it follows the LINE assumptions, the p-values for all the Beta coefficients in this model are significant, less than 0.05, and this model has the highest r_squared (0.904) and adj_r_squared (0.9) values.

Although this model is not perfect because there is a high correlation between the childmort, chdperwoman and water variable we can utilize this model for predictions.

Secondary Final Model: Model for Inferences:

Our best model although it has a high correlation between the independent variables, it also has a good adjusted R^2 value of 0.9. Although this model is good for predicting variables it cannot be used to make inferences due to the high collinearity between the independent variables.

Hence, we also present our best model in terms of making inferences.

InferenceModel <- lm(data = data, lifeexp ~ childmort + water)

We started with combinations of two variables through interaction models and quadratic models. Although all the interaction models we tried did not have a significant value, that is, the pvalue was greater than 0.05 for the interaction term, while the quadratic model did not make a significant difference to the model overall.

As we can see below using an example of a quadratic model, there isn't much difference between the two. The R^2 value remains the same, the significance of p-values remain the same.

```
Modelquad <- lm(data = data, lifeexp ~  childmort + water + water^2)
get_regression_table(Modelquad)
```

```
## # A tibble: 3 × 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   69.9       3.34      20.9    0       63.3     76.5  
## 2 childmort   -0.163     0.015    -11.2    0       -0.192   -0.134
## 3 water        0.091     0.034      2.71   0.008    0.024    0.158
```

```
get_regression_summaries(Modelquad)
```

```
## # A tibble: 1 × 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.845         0.842  14.3  3.77  3.82      327.       0     2   123
```

Hence, after looking at Cor matrix and running different combinations we concluded to the below model.

```
InferenceModel <- lm(data = data, lifeexp ~  childmort + water)
get_regression_table(InferenceModel)
```

```
## # A tibble: 3 × 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   69.9       3.34      20.9    0       63.3     76.5  
## 2 childmort   -0.163     0.015    -11.2    0       -0.192   -0.134
## 3 water        0.091     0.034      2.71   0.008    0.024    0.158
```

```
get_regression_summaries(InferenceModel)
```

```
## # A tibble: 1 × 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.845         0.842  14.3  3.77  3.82      327.       0     2   123
```

How the independent variables affect the dependent variable:

The expected increase in life expectancy when childmort increases by 1 unit and water is held constant is -0.163 => y = 69.9 - 0.163x

The expected increase in life expectancy when water increases by 1 unit and childmort is held constant is 0.091 => y = 69.9 + 0.091x

This model is not perfect as there are some outliers, but out of all the independent variables that were tested, these two variables do not possess multicollinearity, they have significant Pvalues and they provide a high R^2 value of 0.845.

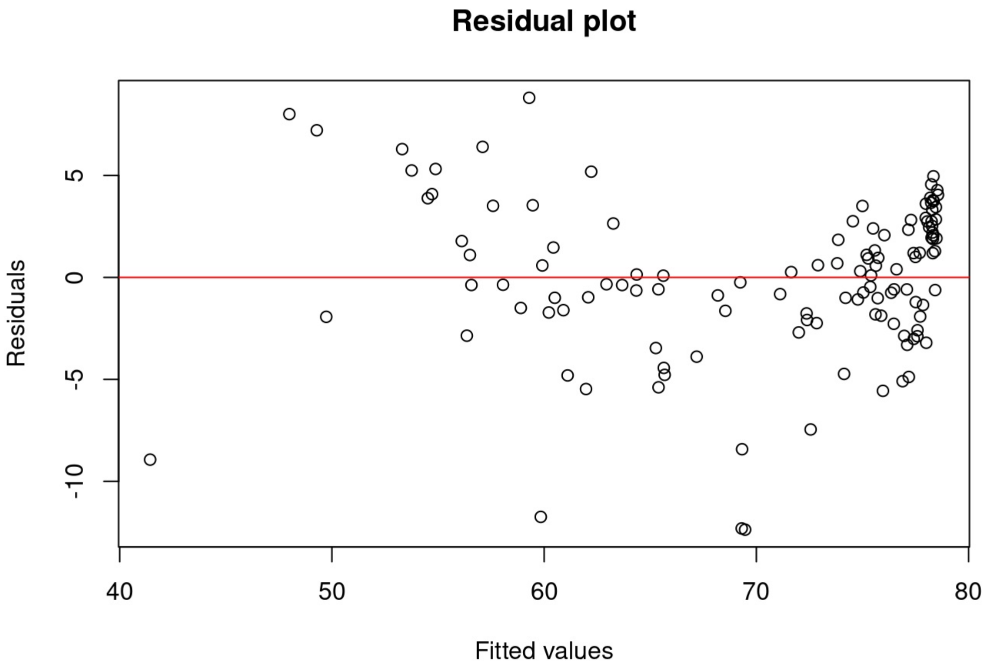
Residual Analysis:

The residual scatterplot has some outliars that cause some LINE assumptions to be violated. It is not uniformly distributed around the mean.

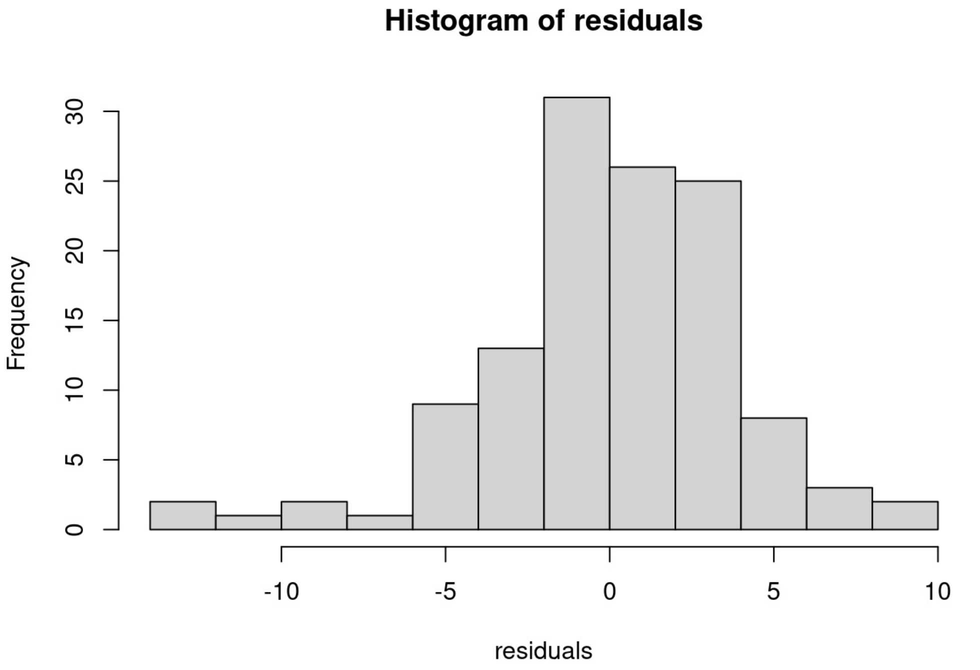
The histogram appears to skewed to the left in the distribution, which indicates that the data has a lower frequency of values on the left side and a higher frequency of values on the right side

The residuals are somewhat normally distributed, the points on the plot mainly fall along a straight line. There are some outliars that deviate from a straight line which indicates slight departures from normality.

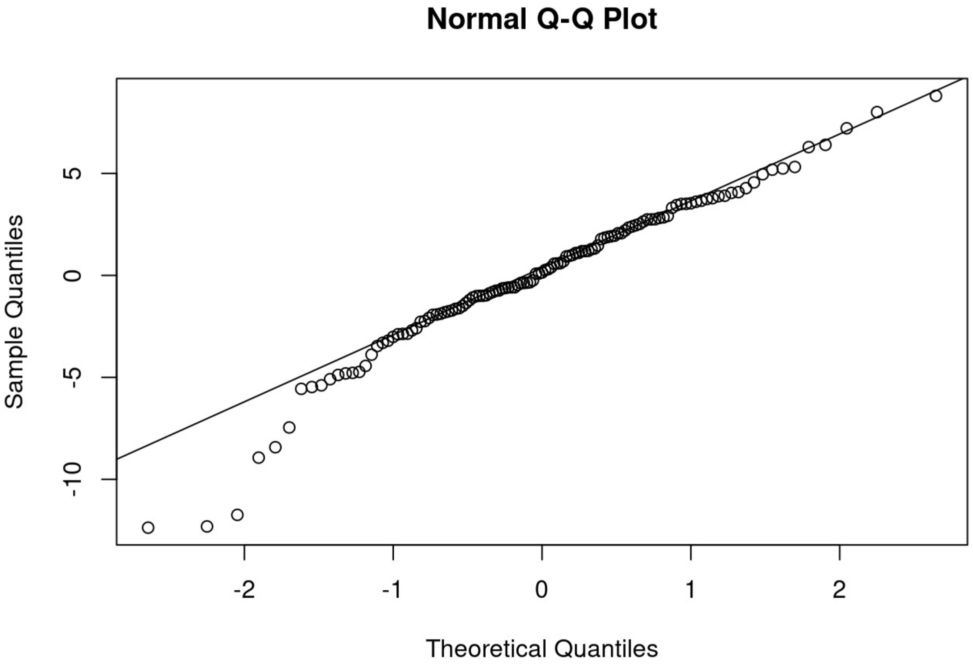
```
model4.1 <- lm(data = data, lifeexp ~ childmort + water)

residuals <- residuals(model4.1)
plot(model4.1$fitted.values, residuals,
     xlab = "Fitted values", ylab = "Residuals",
     main = "Residual plot")

abline(h = 0, col = "red")
```



```
hist(residuals)
```



```
qqnorm(residuals)
qqline(residuals)
```



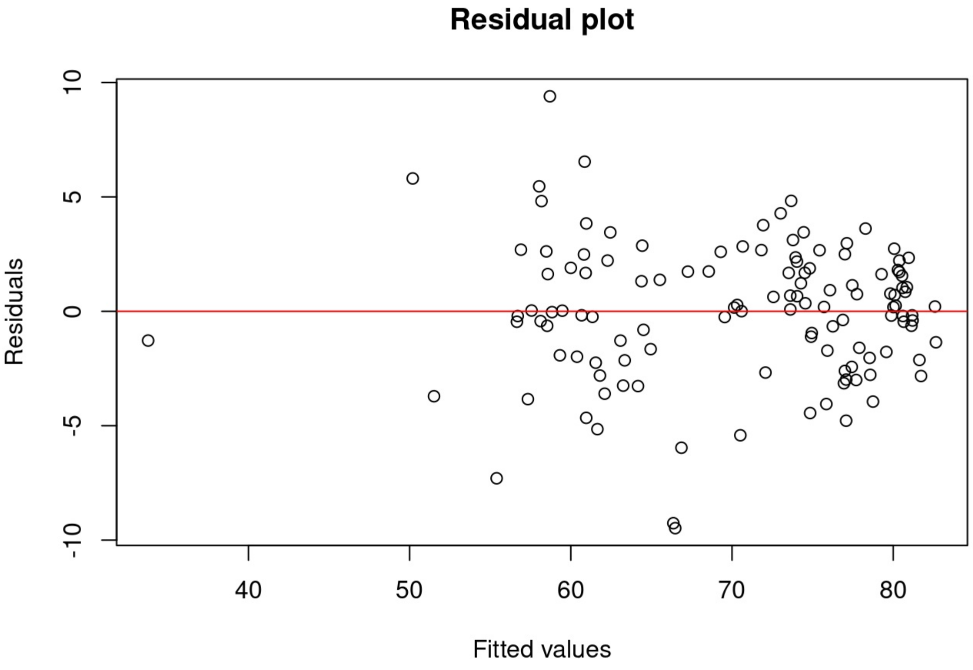
The residuals appear to be randomly scattered around the horizontal line at zero, which indicates that the linearity and constant variance assumptions are likely met. There are no obvious patterns in the residuals that suggest a violation of these assumptions.

The residuals appear to be approximately normally distributed. The histogram is roughly symmetric and bell-shaped, which is a good indication that the normality assumption is met.

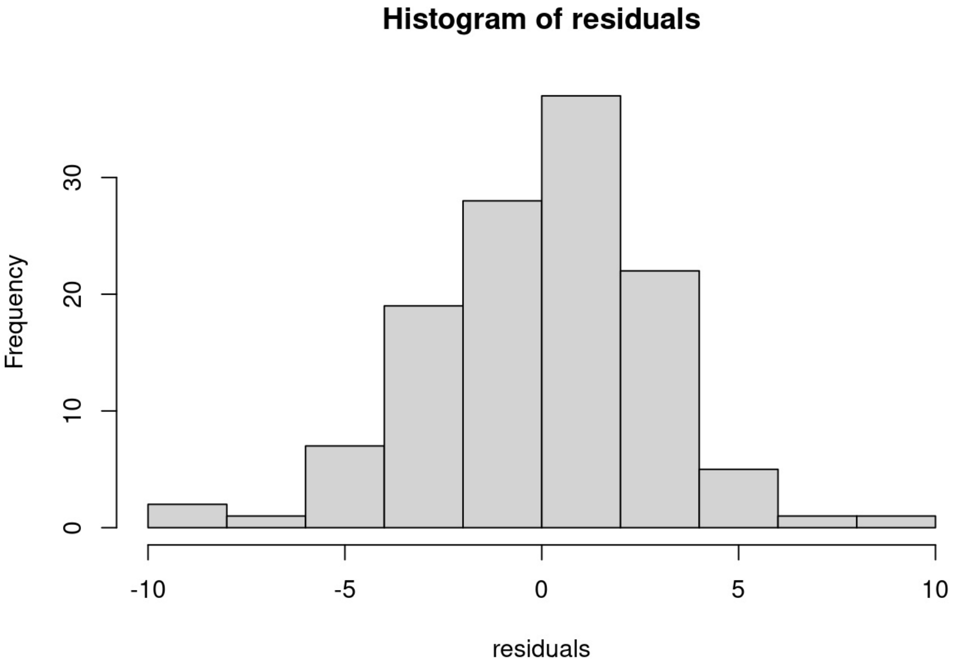
The residuals are normally distributed, the points on the plot mainly fall along a straight line. There are some outliars that deviate from a straight line which indicates slight departures from normality.

Overall, based on the scatterplot and histogram of residuals, it appears that the assumptions of linearity, constant variance, independence, and normality are reasonably well met. However, it’s always a good idea to perform a more formal assessment of these assumptions using additional diagnostic plots and statistical tests.

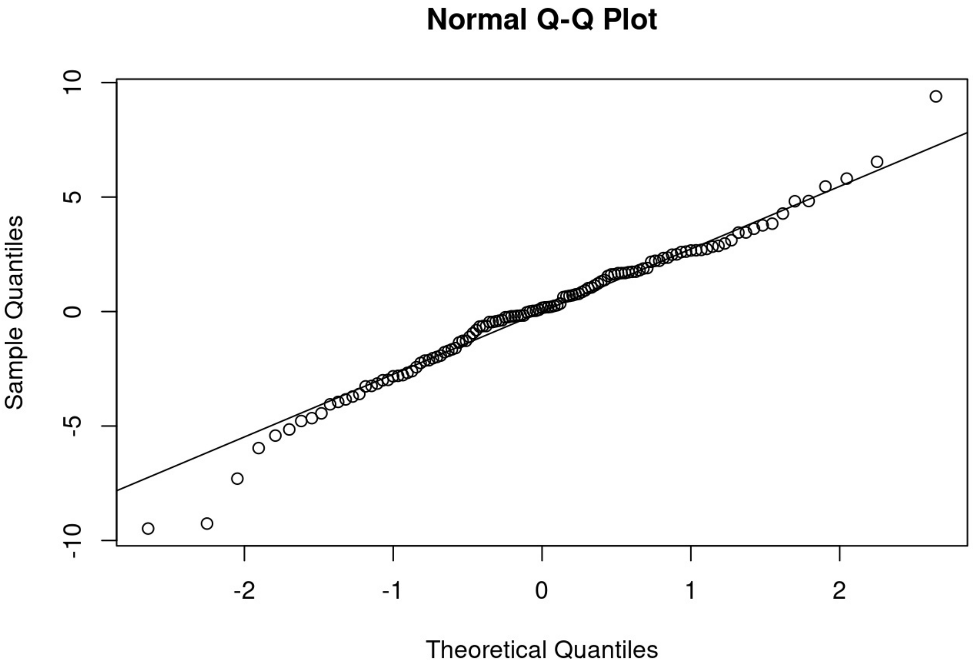
```
model <- lm(data = data, lifeexp ~ healthspend + (chdperwoman * childmort) + water)

residuals <- residuals(model)
plot(model$fitted.values, residuals,
     xlab = "Fitted values", ylab = "Residuals",
     main = "Residual plot")

abline(h = 0, col = "red")
```



```
hist(residuals)
```



```
qqnorm(residuals)
qqline(residuals)
```



Conclusion:

Hence, through exploratory data analysis we made inferences about the independent and dependent variables and their relationships with each other, then through correlation matrix, quadratic and interaction methods we concluded two models. While the Primary model is good for predictions, we can use the secondary model for inferences. Lastly, the residual analysis showed the LINE assumptions of the models.